

Universal compression of Gaussian sources with unknown parameters

Alon Orlitsky[†] and Narayana Santhanam^{*}

October 17, 2014

1

Abstract

For a collection of distributions over a countable support set, the worst case universal compression formulation by Shtarkov attempts to assign a universal distribution over the support set. The formulation aims to ensure that the universal distribution does not underestimate the probability of any element in the support set relative to distributions in the collection. When the alphabet is uncountable and we have a collection \mathcal{P} of Lebesgue continuous measures instead, we ask if there is a corresponding universal probability density function (pdf) that does not underestimate the value of the density function at any point in the support relative to pdfs in \mathcal{P} . An example of such a measure class is the set of all Gaussian distributions whose mean and variance are in a specified range. We quantify the formulation in the uncountable support case with the *attenuation* of the class—a quantity analogous to the worst case redundancy of a collection of distributions over a countable alphabet. An attenuation of A implies that the worst case optimal universal pdf at any point x in the support is always at least the value any pdf in the collection \mathcal{P} assigns to x divided by A . We analyze the attenuation of the worst optimal universal pdf over length- n samples generated *i.i.d.* from a Gaussian distribution whose mean can be anywhere between $-\alpha/2$ to $\alpha/2$ and variance between σ_m^2 and σ_M^2 . We show that this attenuation is finite, grows with the number of samples as $\mathcal{O}(n)$, and also specify the attenuation exactly without approximations. When only one parameter is allowed to vary, we show that the attenuation grows as $\mathcal{O}(\sqrt{n})$, again keeping in line with results from prior literature that fix the order of magnitude as a factor of \sqrt{n} per parameter. In addition, we also specify the attenuation exactly without approximation when only the mean or only the variance is allowed to vary.

Keywords: infinitely divisible distributions, universal compression, uncountable support, Gaussians distributions.

Compression has been well studied since Shannon [1] formalized not just the notion of what it means to represent data or signals in a compact form, but also quantified how compact the representation can be. For data that come from a countable (discrete) alphabet, this lower bound on compression is essentially the entropy of the source. Furthermore, concrete schemes to represent discrete data in bits are also known—the Huffman coding scheme being the optimal one.

While the quantification of the limits of compression is elegant, it does not take into account one of the practicalities of compression—we do not know the underlying distribution. Instead,

[†] Department of Electrical and Computer Engineering, University of California, San Diego
Email: alon@ucsd.edu

^{*} Department of Electrical Engineering, University of Hawaii at Manoa
Email: nsanthan@hawaii.edu

Shtarkov [2] considered a compression framework where the underlying distribution remains unknown, assuming instead that the unknown distribution belongs to a known collection \mathcal{P} of possible distributions. The framework in [2] is a natural approach to *universal compression* [3], where we attempt to describe the data almost as well as the underlying model by means of a *universal* distribution. Suppose we have a class \mathcal{P} of distributions over a countable set X . We now attempt to find a universal distribution q over X such that

$$\sup_{x \in X} \sup_{p \in \mathcal{P}} \frac{p(x)}{q(x)} \tag{1}$$

is as small as possible. The ratio above is always ≥ 1 , since for any two distributions p and q over X

$$\sup_{x \in X} \frac{p(x)}{q(x)} \geq 1.$$

Suppose the supremum in (1) is finite and equals A . It follows that no matter what the realization x is or the underlying model p is

$$q(x) \geq \frac{p(x)}{A}.$$

Therefore, where A is suitably close to 1, the universal q obviates the need to know the underlying distribution p within \mathcal{P} .

When we deal with sequences of variables, the quantity of interest is often not the entire sequence itself. Rather, we may be interested in different statistics of a sequence. For example, in the *i.i.d.* case, the sum of the sequence of random variables is a sufficient statistic.

There are several large deviation results that help us tackle such statistics better. Indeed, large deviation results for sums of many kinds of sequences of random variables are well known. At the very simplest, the sum of *i.i.d.* Bernoulli random variables is distributed as Gaussian in the limit as the number of variables increases to infinity. The mean of the Gaussian equals to the mean of the Bernoulli random variable and its variance is easily related to the Bernoulli variable's variance.

More generally, the limit probability law need not always be Gaussian as above even when we consider the component random variables to be binary. With appropriate scaling of probabilities of the individual binary random variables, we can have the limiting law be Poisson for example [4, vol 2, p173]. Indeed, the different distributions that could come up as the limiting law of sums of random variables are characterized as *infinitely divisible* distributions (see *e.g.* [4, vol 2, ch 6] or [5]). This family of infinitely divisible distributions includes several well known distributions such as the negative binomial, Gamma, χ^2 and Cauchy distributions, in addition to Gaussians and Poisson distributions. In all these cases, it is natural to use the limiting infinitely divisible distribution to describe the sum, rather than the sequence of random variables.

For more general functions other than the sum, deviation bounds such as Hoeffding's inequality or McDiarmid's inequality (among others) allow us to define a dominating distribution on the deviation of the function from its mean value. Often, the mean of these dominating distributions is easily obtained and the general fall off of large deviations. Describing these functions is therefore better handled by describing the dominating distribution rather than the sequence itself.

In both cases—whether we consider infinitely divisible distributions or distributions that characterize large deviations, we may have to deal with a family of distributions with uncountable support such as Gaussian distributions. The exact parameters of the distribution in question is a function of the underlying statistics of the sequence though the family the distribution belongs to is fixed to within the range of parameters. The natural question then is, in analogy with how

we deal with countable data, can we universally handle these collections of distributions over uncountable supports as well?

Let X be an uncountable set, say the real line. Suppose, as before, that \mathcal{P} is the collection of probability measures over X . In addition, the measures in \mathcal{P} are absolutely continuous with respect to the Lebesgue measure. We see data from an unknown measure in \mathcal{P} . Could we take a universal approach again and come up with a universal pdf for \mathcal{P} that does not underestimate the true density anywhere?

Surprisingly, despite the strong motivation, the uncountable support case has received very little attention, despite the seminal work of Rissanen [6]. The multitude of results [7, 8, 9, 10] on universal compression over finite alphabets do not apply non-trivially when the domain is uncountable.

One exception is Rissanen’s results in [6] that indicates that even while the support may be uncountable, if the class \mathcal{P} of probability measures can be parameterized by a few parameters there must be a good universal measure for \mathcal{P} . Formally, we define the attenuation of a collection \mathcal{P} of measures over the support $X \subseteq \mathbb{R}$ in analogy with Shtarkov [2] and Rissanen [6]. Suppose every measure in the collection \mathcal{P} is absolutely continuous with the Lebesgue measure for the sake of simplicity. Then, we define the attenuation

$$\hat{A}(\mathcal{P}) = \inf_q \sup_{x \in X} \sup_{p \in \mathcal{P}} \frac{p(x)}{q(x)},$$

where $p \in \mathcal{P}$ and q are the probability distribution functions (pdfs) with respect to the Lebesgue measure defined in the standard way. We also let for all $x \in X$,

$$\hat{p}(x) = \sup_{p \in \mathcal{P}} p(x).$$

Remark The requirement of absolute continuity with the Lebesgue measure can be relaxed in several ways. One way is to decompose measures into a discrete probability distribution and a probability density function. It is also possible to have a more general (and cleaner, if more abstract) formulation where we simply require all $p \in \mathcal{P}$ to be absolutely continuous with respect to the universal q , and consider the Radon-Nikodym derivative in place of pdf ratios. However, we keep the restriction in this paper to focus on Gaussian probability density functions. \square

To make the problem concrete, we consider collections of Gaussian distributions over the real line. If, as in the Gaussian case, $\hat{p}(x)$ is measurable we clearly have

$$\hat{A}(\mathcal{P}) = \int_X \hat{p}(x) dx.$$

If the integral above is bounded, we say that the *attenuation* of \mathcal{P} is finite. Here the pdf q^* that achieves the infimum in the definition of attenuation above is easily seen to be

$$q^*(x) = \frac{\hat{p}(x)}{\int_{x' \in X} \hat{p}(x') dx'}.$$

In particular, we also consider the case where X is the space of sequences of real numbers (sampled *i.i.d.*) from distributions in \mathcal{P} .

This paper studies collections of Gaussian distributions over real numbers. As mentioned before, Gaussian distributions form the limit law of sums of a wide variety of *i.i.d.* random variables—see [4, 11] for more details. When the individual random variables can be from alphabets other than binary, it is possible to characterize the mean of the limit law without knowing

the variance, and vice versa. Furthermore, we also study the attenuation of sequences of *i.i.d.* Gaussian random variables—corresponding to describing disjoint partial sums of a sequence of an unknown *i.i.d.* random variables.

We consider two cases. In the first case, only one parameter (either the mean or the variance) is unknown while the other is specified. These results appear in Theorem 1 and Corollary 6. The second case allows both the mean and variance to be unknown, and is treated in Theorems 4 and 5. In both cases, we will also calculate the attenuation of *i.i.d.* sampling from the Gaussian collection precisely, without any approximations. These results extend and make more precise a specific section on results on similar collections in [6].

A word on notation—we will use bold font to denote vectors and matrices. The transpose of a matrix \mathbf{K} is \mathbf{K}^T and its determinant is $|\mathbf{K}|$. For a vector $\mathbf{x} = (x_1, \dots, x_n)$, we use $d\mathbf{x}$ to denote $dx_1 \dots dx_n$. We will interchangeably refer to length- n sequences x_1, \dots, x_n by their length- n column vector analogs $\mathbf{x} = (x_1, \dots, x_n)^T$.

1 Gaussians with unknown mean and variance 1

Let G_α be the collection of gaussians with variance $\sigma^2 = 1$ and where the mean μ lies in the range $-\alpha/2 \leq \mu \leq \alpha/2$ (total range is α). We denote by G_α^n the collection of all pdfs on \mathbb{R}^n obtained by *i.i.d.* sampling from a distribution in G_α . We will do a couple of examples before computing the attenuation for length- n strings from the class G_α^n for general n . In this section, for any length- n sequence \mathbf{x} , we denote $\hat{p}(\mathbf{x}) = \arg \max_{p \in G_\alpha^n} p(\mathbf{x})$.

Example 1. (Length 1) If $-\alpha/2 \leq x \leq \alpha/2$, the Gaussian in G_α maximizing $p(x)$ has mean x , hence $\hat{p}(x) = \frac{1}{\sqrt{2\pi}}$. If $x \geq \alpha/2$, the Gaussian in G_α maximizing $p(x)$ has mean $\alpha/2$, hence $\hat{p}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\alpha/2)^2}$. Similarly for $x \leq -\alpha/2$.

The attenuation for a sequence of length 1 is therefore

$$\begin{aligned} \hat{A}(G_\alpha^1) &= \int_{-\infty}^{\infty} \hat{p}(x) dx \\ &= \int_{-\infty}^{-\alpha/2} \hat{p}(x) dx + \int_{-\alpha/2}^{\alpha/2} \hat{p}(x) dx + \int_{\alpha/2}^{\infty} \hat{p}(x) dx \\ &= \frac{1}{2} + \frac{\alpha}{\sqrt{2\pi}} + \frac{1}{2} \\ &= 1 + \frac{\alpha}{\sqrt{2\pi}}, \end{aligned}$$

which makes sense as if $\alpha = 0$, we know the distribution and have attenuation 1. □

Next consider attenuation for sequences of length 2.

Example 2. Let x_1 and x_2 denote the first and second outcomes. Define $y = (x_1 + x_2)/2$ to be the average and $z = x_1 - y = (x_1 - x_2)/2$ to be the difference between x_1 and the average. A gaussian with mean μ will assign the sequence (x_1, x_2) probability

$$p(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}[(x_1-\mu)^2 + (x_2-\mu)^2]}$$

Therefore the maximum likelihood (ML) Gaussian in G_α^2 , \hat{p} , has mean $\mu = y$ if $-\alpha/2 \leq y \leq \alpha/2$, has $\mu = \alpha/2$ if $y > \alpha/2$, and $\mu = -\alpha/2$ if $y < -\alpha/2$.

It follows that the attenuation for 2-element sequences is

$$\begin{aligned}
\hat{A}(G_\alpha^2) &= \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \hat{p}(x_1, x_2) \\
&= 2 \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dz \hat{p}(y, z) \\
&= \frac{2}{(\sqrt{2\pi})^2} \int_{-\alpha/2}^{\alpha/2} dy \int_{-\infty}^{\infty} dz \exp\left(-\frac{1}{2}z^2 - \frac{1}{2}z^2\right) \\
&\quad + \frac{2 \cdot 2}{(\sqrt{2\pi})^2} \int_{\alpha}^{\infty} dy \int_{-\infty}^{\infty} dz \exp\left(-\frac{1}{2}\left(\left(\frac{\alpha}{2} - (y - z)\right)^2 + \left(\frac{\alpha}{2} - (y + z)\right)^2\right)\right).
\end{aligned}$$

Now, the first summand is

$$\frac{\sqrt{2}}{\sqrt{2\pi}} \int_{-\alpha/2}^{\alpha/2} dy \frac{\sqrt{2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dz \exp(-z^2) = \frac{\sqrt{2}}{\sqrt{2\pi}} \int_{-\alpha/2}^{\alpha/2} dy = \frac{\alpha}{\sqrt{\pi}},$$

and the second summand is

$$\frac{2\sqrt{2}}{\sqrt{2\pi}} \cdot \int_{\alpha/2}^{\infty} dy \frac{\sqrt{2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dz \exp\left(-\left(y - \frac{\alpha}{2}\right)^2 - z^2\right) = 1.$$

So

$$\hat{A}(G_\alpha^2) = 1 + \frac{\alpha}{\sqrt{\pi}}.$$

□

Theorem 1. For all n ,

$$\hat{A}(G_\alpha^n) = 1 + \alpha \sqrt{\frac{n}{2\pi}}.$$

Proof The case $n = 1$ has been considered in Example 1. For $n \geq 2$, we will transform the length n sequence $\mathbf{x} = (x_1, \dots, x_n)^T$ into the following variables

$$y = (x_1 + \dots + x_n)/n, \text{ and } z_j = x_j - y \text{ for } 1 \leq j \leq n-1.$$

Now $\mathbf{z} = (z_1, \dots, z_{n-1})^T$ takes values in \mathbb{R}^{n-1} and $y \in \mathbb{R}$. Then the Jacobian of the transformation,

$$\frac{\partial y \mathbf{z}}{\partial \mathbf{x}} = \frac{1}{n} \begin{pmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ n-1 & -1 & -1 & \dots & -1 & -1 \\ -1 & n-1 & -1 & \dots & -1 & -1 \\ -1 & -1 & n-1 & & -1 & -1 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ -1 & -1 & -1 & \dots & n-1 & -1 \end{pmatrix},$$

and its determinant,

$$\left| \frac{\partial y \mathbf{z}}{\partial \mathbf{x}} \right| = (-1)^{n-1} \frac{((n-1)+1)^{n-1}}{n^n} = \frac{(-1)^{n-1}}{n}. \quad (2)$$

We will compute the attenuation using the above transformation. The length- n attenuation,

$$\begin{aligned}
\hat{A}(G_\alpha^n) &= \int \hat{p}(\overline{\mathbf{x}}) d\mathbf{x} \\
&= n \int \hat{p}(y, z_1, \dots, z_{n-1}) dy d\mathbf{z} \\
&= 2 \frac{n}{(\sqrt{2\pi})^n} \int_{y=\alpha/2}^{\infty} \int_{z_1, \dots, z_{n-1}=-\infty}^{\infty} \exp\left(-\frac{\sum_{i=1}^{n-1} \left(\frac{\alpha}{2} - (y + z_i)\right)^2 + \left(\frac{\alpha}{2} - (y - \sum_{i=1}^{n-1} z_i)\right)^2}{2}\right) dy d\mathbf{z} \\
&\quad + \frac{n}{(\sqrt{2\pi})^n} \int_{y=-\alpha/2}^{\alpha/2} \int_{z_1, \dots, z_{n-1}=-\infty}^{\infty} \exp\left(-\frac{\sum_{i=1}^{n-1} z_i^2 + \left(\sum_{i=1}^{n-1} z_i\right)^2}{2}\right) dy d\mathbf{z}.
\end{aligned}$$

We simplify the first integral in the last line above using

$$\sum_{i=1}^{n-1} \left(\frac{\alpha}{2} - (y + z_i)\right)^2 + \left(\frac{\alpha}{2} - (y - \sum_{i=1}^{n-1} z_i)\right)^2 = n\left(\frac{\alpha}{2} - y\right)^2 + \sum_{i=1}^{n-1} z_i^2 + \left(\sum_{i=1}^{n-1} z_i\right)^2.$$

Doing so, and letting

$$I \stackrel{\text{def}}{=} \frac{\sqrt{n}}{(\sqrt{2\pi})^{n-1}} \int_{z_1, \dots, z_{n-1}=-\infty}^{\infty} \exp\left(-\frac{\sum_{i=1}^{n-1} z_i^2 + \left(\sum_{i=1}^{n-1} z_i\right)^2}{2}\right) d\mathbf{z},$$

we obtain

$$\hat{A}(G_\alpha^n) = I \cdot \left(2 \frac{\sqrt{n}}{\sqrt{2\pi}} \int_{y=\alpha/2}^{\infty} \exp\left(-\frac{1}{2} \left(n\left(\frac{\alpha}{2} - y\right)^2\right)\right) dy + \frac{\sqrt{n}\alpha}{\sqrt{2\pi}} \right) = I \cdot \left(1 + \alpha \sqrt{\frac{n}{2\pi}} \right).$$

We will now show that the integral $I = 1$ to conclude the proof of the theorem. Write

$$\sum_{i=1}^{n-1} z_i^2 + \left(\sum_{i=1}^{n-1} z_i\right)^2 = \sum_{i=1}^{n-1} 2z_i^2 + \sum_{1 \leq i, j \leq n-1} z_i z_j = \mathbf{z}^T \mathbf{K}^{-1} \mathbf{z} \quad (3)$$

where \mathbf{K}^{-1} is a $(n-1) \times (n-1)$ matrix, given by

$$\mathbf{K}^{-1} = \begin{pmatrix} 2 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ 1 & 1 & 2 & & 1 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 2 \end{pmatrix}.$$

Now letting \mathbf{I}_{n-1} be an identity matrix of dimensions $(n-1) \times (n-1)$ and writing $\mathbf{1}$ for a column vector of $n-1$ ones, we have

$$\mathbf{K}^{-1} = \mathbf{I}_{n-1} + \mathbf{1}\mathbf{1}^T.$$

It follows from the Syvelster determinant theorem [12] that

$$|\mathbf{K}^{-1}| = |\mathbf{I}_{n-1} + \mathbf{1}\mathbf{1}^T| = |1 + \mathbf{1}^T \mathbf{1}| = 1 + (n-1) = n. \quad (4)$$

Hence

$$|\mathbf{K}| = \frac{1}{n}.$$

Therefore,

$$\begin{aligned} I &= \frac{\sqrt{n}}{(\sqrt{2\pi})^{n-1}} \int_{z_1, \dots, z_{n-1} = -\infty}^{\infty} \exp \left(-\frac{\sum_{i=1}^{n-1} z_i^2 + \left(\sum_{i=1}^{n-1} z_i \right)^2}{2} \right) dz \\ &= \frac{1}{(\sqrt{2\pi})^{n-1} \sqrt{|\mathbf{K}|}} \int_{z_1, \dots, z_{n-1} = -\infty}^{\infty} \exp \left(-\frac{\bar{z}^T \mathbf{K}^{-1} \bar{z}}{2} \right) dz \\ &= 1. \end{aligned} \tag{5}$$

The theorem follows. \square

2 Gaussians with unknown mean and variance

Let $G_{\alpha, \sigma_m, \sigma_M}$ be the collection of iid gaussians with $-\alpha/2 \leq \mu \leq \alpha/2$ and $\sigma_m \leq \sigma \leq \sigma_M$. Throughout this section, we will use

$$p_{\sigma, \mu}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(x - \mu)^2}{2\sigma^2} \right)$$

to denote a Gaussian with variance σ^2 and mean μ . As before, we denote the collection of all pdfs on \mathbb{R}^n obtained by *i.i.d.* sampling from a distribution in $G_{\alpha, \sigma_m, \sigma_M}$ by $G_{\alpha, \sigma_m, \sigma_M}^n$. As before, in this section, for any length- n sequence \mathbf{x} , we denote $\hat{p}(\mathbf{x}) = \arg \max_{p \in G_{\alpha, \sigma_m, \sigma_M}^n} p(\mathbf{x})$.

Example 3. Let

$$p_{\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{x^2}{2\sigma^2} \right)$$

denote the Gaussian distribution with zero mean and standard deviation σ . Differentiating $\log p_{\sigma}(x)$ with respect to σ , we obtain that for every x , $p_{\sigma}(x)$ is maximized by $\sigma = |x|$. Therefore

$$\max_{\sigma} p_{\sigma}(x) = \frac{1}{\sqrt{2\pi}x} \exp \left(-\frac{1}{2} \right) = \frac{1}{\sqrt{2\pi}ex}.$$

It follows that

$$\hat{p}(x) = \begin{cases} \frac{1}{\sqrt{2\pi} \cdot \sigma_m} & 0 \leq |x| \leq \frac{\alpha}{2} \\ \frac{1}{\sqrt{2\pi} \cdot \sigma_m} \exp \left(-\frac{1}{2} \frac{(|x| - \alpha/2)^2}{\sigma_m^2} \right) & \frac{\alpha}{2} \leq |x| \leq \frac{\alpha}{2} + \sigma_m \\ \frac{1}{\sqrt{2\pi}e \cdot (|x| - \alpha/2)} & \frac{\alpha}{2} + \sigma_m \leq |x| \leq \frac{\alpha}{2} + \sigma_M \\ \frac{1}{\sqrt{2\pi} \cdot \sigma_M} \exp \left(-\frac{1}{2} \frac{(|x| - \alpha/2)^2}{\sigma_M^2} \right) & \frac{\alpha}{2} + \sigma_M \leq |x|. \end{cases}$$

Hence

$$\hat{A}(G_{\alpha, \sigma_m, \sigma_M}^1) = 1 + \frac{\alpha}{\sigma_m} \cdot \sqrt{\frac{1}{2\pi}} + \sqrt{\frac{2}{\pi e}} \cdot \ln \frac{\sigma_M}{\sigma_m}. \quad \square$$

Given a sequence $\mathbf{x} = (x_1, \dots, x_n)^T$, we let as before

$$y = \frac{\sum_{i=1}^n x_i}{n}.$$

Furthermore for a_- , a_+ and A in \mathbb{R} , let

$$(A)_{a_-}^{a_+} = \begin{cases} a_- & A \leq a_- \\ A & a_- \leq A \leq a_+ \\ a_+ & a_+ \leq A. \end{cases}$$

The following lemma characterizes the maximum likelihood distribution.

Lemma 2. For $x_1, \dots, x_n \in \mathbb{R}^n$, the ML estimates of the mean and variance are

$$\hat{\mu} = (y)_{-\alpha/2}^{\alpha/2}$$

and

$$\hat{\sigma}^2 = \left(\frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{n} \right)_{\sigma_m^2}^{\sigma_M^2}.$$

Namely,

$$\arg \max_{\sigma, \mu} p_{\sigma, \mu}(x_1, \dots, x_n) = p_{\hat{\sigma}, \hat{\mu}}.$$

Proof If

$$(y)_{-\alpha/2}^{\alpha/2} = y \text{ and } \left(\frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{n} \right)_{\sigma_m^2}^{\sigma_M^2} = \frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{n}$$

the lemma follows by noting that the first partial derivatives of $p_{\sigma, \mu}$ are zero at $\hat{\mu} = y$ and $\hat{\sigma} = \frac{\sum_{i=1}^n (x_i - y)^2}{n}$. The second partial derivatives can be easily verified to satisfy

$$\frac{\partial^2 p_{\sigma, \mu}}{\partial \sigma \partial \mu} - \frac{\partial^2 p_{\sigma, \mu}}{\partial \sigma^2} \frac{\partial^2 p_{\sigma, \mu}}{\partial \mu^2} < 0$$

with $\frac{\partial^2 p_{\sigma, \mu}}{\partial \sigma^2} < 0$ and $\frac{\partial^2 p_{\sigma, \mu}}{\partial \mu^2} < 0$, meeting the conditions for a maxima of $p_{\sigma, \mu}$. If $(y)_{-\alpha/2}^{\alpha/2} \neq y$, the corresponding first derivative of $p_{\sigma, \mu}$ at $\mu = \hat{\mu}$ is non-zero. Therefore, moving to the interior of the parameter space parallel to the direction of a unit vector along μ decreases $p_{\sigma, \mu}$. Hence $p_{\sigma, \mu}$ must be maximized on the boundary. A similar observation holds for σ as well. The lemma follows.

Finally, we will need Stirling's approximation of the Gamma function.

Lemma 3. (Stirling) $\Gamma(x+1) = \sqrt{2\pi x} \left(\frac{x}{e}\right)^x (1 + \mathcal{O}(\frac{1}{x}))$. □

We are now in a position to compute the attenuation of $G_{\alpha, \sigma_m, \sigma_M}^N$. The main theorem below, Theorem 4 uses the Stirling's approximation to simplify results into a easily readable form. Theorem 5 gives the same result in precise form, though it is unwieldy.

Theorem 4. For $n \geq 2$,

$$\hat{A}(G_{\alpha, \sigma_m, \sigma_M}^n) = \frac{\alpha \sqrt{n(n-1)}}{\pi \sqrt{2}} \left(\frac{1}{\sigma_m} - \frac{1}{\sigma_M} \right) + \frac{\alpha \sqrt{n}}{\sqrt{2\pi}} \left(\frac{I_n}{\sigma_m} + \frac{1 - I_n}{\sigma_M} \right) + \sqrt{\frac{n}{\pi}} \ln \frac{\sigma_M}{\sigma_m} + \mathcal{O}(1),$$

where

$$I_n \stackrel{\text{def}}{=} \frac{\sqrt{n}}{(\sqrt{2\pi})^{n-1}} \int_{\sum_{i=1}^{n-1} z_i^2 + (\sum_{j=1}^{n-1} z_j)^2 \leq n} \exp\left(-\frac{\sum_{i=1}^{n-1} z_i^2 + \left(\sum_{i=1}^{n-1} z_i\right)^2}{2}\right) dz.$$

As $n \rightarrow \infty$, we have $I_n \rightarrow 1$.

Proof Denote $\mathbf{x} = (x_1, \dots, x_n)^T$. We compute the integral

$$\int_{\mathbf{x}} \hat{p}(\mathbf{x}) d\mathbf{x}$$

by splitting the domain of the integral, first based on the value of the mean followed by the value of the ML estimate of the variance. Specifically, we partition

$$\mathbb{R}^n = \mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3,$$

where $\mathcal{R}_1 \stackrel{\text{def}}{=} \{\mathbf{x} : \mathbf{x}^T \mathbf{1} \leq -n\alpha/2\}$, $\mathcal{R}_2 \stackrel{\text{def}}{=} \{\mathbf{x} : -n\alpha/2 \leq \mathbf{x}^T \mathbf{1} \leq n\alpha/2\}$, and $\mathcal{R}_3 \stackrel{\text{def}}{=} \{\mathbf{x} : \mathbf{x}^T \mathbf{1} \geq n\alpha/2\}$. We consider each of the regions separately below. Regions \mathcal{R}_1 and \mathcal{R}_3 contribute $\mathcal{O}(\sqrt{n})$ terms each, while \mathcal{R}_2 contributes $\mathcal{O}(n)$. This is to be expected since both parameters are in play in \mathcal{R}_2 while only one (the variance) is effectively in play in \mathcal{R}_1 and \mathcal{R}_3 .

Region \mathcal{R}_1 In this region, the ML estimate of the mean is $-\alpha/2$. Depending on the ML estimate of the variance, we further subdivide

$$\mathcal{R}_1 = \mathcal{R}_{11} \cup \mathcal{R}_{12} \cup \mathcal{R}_{13},$$

where

$$\begin{aligned} \mathcal{R}_{11} &\stackrel{\text{def}}{=} \{\mathbf{x} \in \mathcal{R}_1 : (\mathbf{x} + \alpha/2)^T (\mathbf{x} + \alpha/2) \leq n\sigma_m^2\}, \\ \mathcal{R}_{12} &\stackrel{\text{def}}{=} \{\mathbf{x} \in \mathcal{R}_1 : n\sigma_m^2 \leq (\mathbf{x} + \alpha/2)^T (\mathbf{x} + \alpha/2) \leq n\sigma_M^2\}, \\ \mathcal{R}_{13} &\stackrel{\text{def}}{=} \{\mathbf{x} \in \mathcal{R}_1 : (\mathbf{x} + \alpha/2)^T (\mathbf{x} + \alpha/2) \geq n\sigma_M^2\}. \end{aligned}$$

From Lemma 2, we have

$$\begin{aligned} &\int_{\mathcal{R}_{11}} \hat{p}(\mathbf{x}) d\mathbf{x} + \int_{\mathcal{R}_{13}} \hat{p}(\mathbf{x}) d\mathbf{x} \\ &= \frac{1}{(2\pi)^{n/2} \sigma_m^2} \int_{\mathbf{x} \in \mathcal{R}_{11}} \exp\left(-\frac{\sum_{i=1}^n (x_i + \alpha/2)^2}{2\sigma_m^2}\right) d\mathbf{x} \\ &\quad + \frac{1}{(2\pi)^{n/2} \sigma_M^2} \int_{\mathbf{x} \in \mathcal{R}_{13}} \exp\left(-\frac{\sum_{i=1}^n (x_i + \alpha/2)^2}{2\sigma_M^2}\right) d\mathbf{x} \\ &\stackrel{(a)}{=} \frac{1}{(2\pi)^{n/2}} \int_{\substack{\mathbf{u} \in \mathbb{R}^n \\ \mathbf{u}^T \mathbf{1} \geq 0}} \exp\left(-\frac{\mathbf{u}^T \mathbf{u}}{2}\right) d\mathbf{u} \\ &= \frac{1}{2}. \end{aligned}$$

In the above, we obtain (a) by transforming the variables in the first integral using $u_i = (x_i + \alpha/2)/\sigma_m$ and the second integral using $u_i = (x_i + \alpha/2)/\sigma_M$. Note as before that $\mathbf{u} = (u_1, \dots, u_n)^T$.

Meanwhile,

$$\begin{aligned}
\int_{\mathcal{R}_{12}} \hat{p}(\mathbf{x}) d\mathbf{x} &\stackrel{(a)}{=} \int_{\sigma_m^2 \leq \frac{\mathbf{u}^T \mathbf{1}}{n} \leq \sigma_M^2} \frac{n^{n/2} e^{-n/2} d\mathbf{u}}{(2\pi)^{n/2} (\mathbf{u}^T \mathbf{u})^{n/2}} \\
&\stackrel{(b)}{=} \frac{n^{n/2} e^{-n/2}}{(2\pi)^{n/2}} \frac{n\pi^{n/2}}{2\Gamma(\frac{n}{2} + 1)} \int_{r=\sqrt{n}\sigma_m}^{\sqrt{n}\sigma_M} \frac{1}{r} dr \\
&= \frac{1}{2} \sqrt{\frac{n}{\pi}} \ln \frac{\sigma_M}{\sigma_m} + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \\
&\sim \frac{1}{2} \sqrt{\frac{n}{\pi}} \ln \frac{\sigma_M}{\sigma_m}.
\end{aligned} \tag{6}$$

In the above, we get (a) by transforming $u_i = x_i + \alpha/2$. To see (b), we transform \mathbf{u} into polar coordinates and note (e.g. [13]) that the surface area of a n -dimensional unit sphere is

$$\frac{n\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)},$$

while the surface area corresponding to $\mathbf{u}^T \mathbf{1} \geq 0$ is exactly half the above quantity. The next equality follows because Stirling's approximation for the Gamma function above has a multiplicative accuracy of $(1 + \mathcal{O}(\frac{1}{n}))$ as in Lemma 3. The net contribution to the attenuation from region \mathcal{R}_1 is therefore

$$\frac{1}{2} + \frac{1}{2} \sqrt{\frac{n}{\pi}} \ln \frac{\sigma_M}{\sigma_m} + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

Region \mathcal{R}_3 This region contributes an identical amount as \mathcal{R}_1 above.

Region \mathcal{R}_2 To tackle this region, we will need the auxillary variable

$$y = \frac{\sum_{i=1}^n x_i}{n},$$

while we will also define auxillary variables very similar to z_j from Theorem 1. Once again, we partition

$$\mathcal{R}_2 = \mathcal{R}_{21} \cup \mathcal{R}_{22} \cup \mathcal{R}_{23},$$

with

$$\begin{aligned}
\mathcal{R}_{21} &= \{\mathbf{x} \in \mathcal{R}_2 : \sum_{i=1}^n (x_i - y)^2 \leq n\sigma_m^2\}, \\
\mathcal{R}_{22} &= \{\mathbf{x} \in \mathcal{R}_2 : n\sigma_m^2 \leq \sum_{i=1}^n (x_i - y)^2 \leq n\sigma_M^2\}, \\
\mathcal{R}_{23} &= \{\mathbf{x} \in \mathcal{R}_2 : n\sigma_M^2 \leq \sum_{i=1}^n (x_i - y)^2\}.
\end{aligned}$$

We will first consider the regions \mathcal{R}_{21} and \mathcal{R}_{23} . We will focus on the case $n \geq 2$ here since the case $n = 1$ has already been handled by Example 3. The contribution to the attenuation of regions \mathcal{R}_{21} and \mathcal{R}_{23} is

$$\int_{\mathbf{x} \in \mathcal{R}_{21}} \frac{1}{(2\pi)^{n/2} \sigma_m^n} \exp\left(-\frac{\sum_{i=1}^n (x_i - y)^2}{2\sigma_m^2}\right) d\mathbf{x} + \int_{\mathbf{x} \in \mathcal{R}_{23}} \frac{1}{(2\pi)^{n/2} \sigma_M^n} \exp\left(-\frac{\sum_{i=1}^n (x_i - y)^2}{2\sigma_M^2}\right) d\mathbf{x}.$$

For $n \geq 2$ we transform the first integral above corresponding to the contribution of \mathcal{R}_{21} from variables \mathbf{x} to

$$y = \frac{\sum_{i=1}^n x_i}{n}, \text{ and } z_j = \frac{x_j - y}{\sigma_m} \text{ for } 1 \leq j \leq n-1,$$

with the new variable y running from $-\alpha/2$ to $\alpha/2$, and the variables z_1, \dots, z_{n-1} taking all possible values such that $\sum_{j=1}^{n-1} z_j^2 + \left(\sum_{i=1}^{n-1} z_i\right)^2 \leq n$. The Jacobian in this case is computed similar to (2),

$$\left| \frac{\partial y \mathbf{z}}{\partial \mathbf{x}} \right| = \frac{(-1)^n}{n \sigma_m^{n-1}}.$$

The second integral corresponding to the contribution of the region \mathcal{R}_{23} is similarly transformed with variables (please note the reuse of notation z_j for later simplicity)

$$y = \frac{\sum_{i=1}^n x_i}{n}, \text{ and } z_j = \frac{x_j - y}{\sigma_M} \text{ for } 1 \leq j \leq n-1.$$

Recalling from (5) that

$$\frac{\sqrt{n}}{(\sqrt{2\pi})^{n-1}} \int_{z_1, \dots, z_{n-1}} \exp\left(-\frac{\sum_{i=1}^{n-1} z_i^2 + \left(\sum_{i=1}^{n-1} z_i\right)^2}{2}\right) d\mathbf{z} = 1,$$

we obtain that \mathcal{R}_{11} and \mathcal{R}_{13} together contribute

$$\alpha \sqrt{\frac{n}{2\pi}} \left(\frac{I_n}{\sigma_m} + \frac{(1 - I_n)}{\sigma_M} \right)$$

where for $n \geq 2$

$$I_n \stackrel{\text{def}}{=} \frac{\sqrt{n}}{(\sqrt{2\pi})^{n-1}} \int_{\sum_{i=1}^{n-1} z_i^2 + (\sum_{j=1}^{n-1} z_j)^2 \leq n} d\mathbf{z} \exp\left(-\frac{\sum_{i=1}^{n-1} z_i^2 + \left(\sum_{i=1}^{n-1} z_i\right)^2}{2}\right).$$

The case $n = 1$ has already been handled by Example 3. When $n = 1$, we do not have the variables \mathbf{z} as in the definition above. Instead we will define $I_1 = 1$ for consistency with Example 3.

The dominant contribution to the attenuation comes from \mathcal{R}_{22} . This region contributes

$$\int_{\mathcal{R}_{22}} \frac{n^{n/2} e^{-n/2}}{(2\pi)^{n/2} (\sum_{i=1}^n (x_i - y)^2)^{n/2}} d\mathbf{x}.$$

Note that in \mathcal{R}_{22} , $\sum_{i=1}^n (x_i - y)^2 \geq n \sigma_m^2 > 0$. It is also interesting to note that this region is non-existent when $n = 1$. For $n \geq 2$, we begin as in Theorem 1, transforming \mathbf{x} into

$$y = \frac{\sum_{i=1}^n x_i}{n}, \text{ and } z_j = x_j - y \text{ for } 1 \leq j \leq n-1.$$

We then have

$$\begin{aligned}
& \int_{\mathcal{R}_{22}} \frac{n^{n/2} e^{-n/2}}{(2\pi)^{n/2} (\sum_{i=1}^n (x_i - y)^2)^{n/2}} d\mathbf{x} \\
& \stackrel{(a)}{=} \int_{\substack{y, \mathbf{z} \\ -\alpha/2 \leq y \leq \alpha/2 \\ \sigma_m^2 \leq \frac{\mathbf{z}^T \mathbf{K}^{-1} \mathbf{z}}{n} \leq \sigma_M^2}} \frac{n^{n/2} e^{-n/2}}{(2\pi)^{n/2} (\mathbf{z}^T \mathbf{K}^{-1} \mathbf{z})^{n/2}} n dy d\mathbf{z} \\
& \stackrel{(b)}{=} \int_{\substack{y, \mathbf{w} \\ -\alpha/2 \leq y \leq \alpha/2 \\ \sigma_m^2 \leq \mathbf{w}^T \mathbf{w} / n \leq \sigma_M^2}} \frac{n^{n/2} e^{-n/2}}{(2\pi)^{n/2} (\mathbf{w}^T \mathbf{w})^{n/2}} \sqrt{n} dy d\mathbf{w} \\
& = \alpha \int_{\sigma_m^2 \leq \mathbf{w}^T \mathbf{w} / n \leq \sigma_M^2} \frac{n^{n/2} e^{-n/2}}{(2\pi)^{n/2} (\mathbf{w}^T \mathbf{w})^{n/2}} \sqrt{n} dy d\mathbf{w} \\
& \stackrel{(c)}{=} \alpha \frac{n^{n/2} e^{-n/2} S(n-1)}{(2\pi)^{n/2}} \int_{r=\sqrt{n}\sigma_m}^{\sqrt{n}\sigma_M} \frac{\sqrt{n}}{r^2} dr \\
& = \frac{\alpha \sqrt{n(n-1)}}{\pi \sqrt{2}} \left(\frac{1}{\sigma_m} - \frac{1}{\sigma_M} \right) + \mathcal{O}(1). \tag{7}
\end{aligned}$$

Here (a) follows from (3) and because $|\mathbf{K}^{-1}| = n$ from (4). To define \mathbf{w} in (b) first note that (3) implies that K^{-1} is positive definite. We let the Cholesky decomposition of $\mathbf{K}^{-1} = \mathbf{C}^T \mathbf{C}$, and set $\mathbf{w} = \mathbf{C}\mathbf{z}$. From (4) we have the determinant $|C| = \sqrt{n}$ to account for the transformation of variables \mathbf{z} to \mathbf{w} . The equality (c) follows from a transformation of \mathbf{w} into polar coordinates in $n-1$ dimensions, where $S(n-1)$ is the surface area of a sphere in $n-1$ dimensions and is equal to

$$\frac{(n-1)\pi^{\frac{n-1}{2}}}{\Gamma(\frac{n-1}{2} + 1)}.$$

The last line above uses the Stirling approximation which has a multiplicative approximation of $1 + \mathcal{O}(\frac{1}{n})$ as specified in Lemma 3, as well as the approximation $\left(1 + \frac{1}{n-1}\right)^{\frac{n-1}{2}} = \sqrt{e} + \mathcal{O}(\frac{1}{n})$. Therefore we have the $\mathcal{O}(1)$ correction term in the last line. The Theorem follows. \square

If we had not approximated for the Gamma function in the above proof, we would have the precise form for attenuation. The following Theorem 5 does exactly that—it proceeds just as Theorem 4 but leaves steps (6) and (7) as they are without approximations. In addition, Theorem 5 subsumes Example 3 as well.

Theorem 5. For all $n \geq 1$,

$$\hat{A}(G_{\alpha, \sigma_m, \sigma_M}^n) = \frac{\alpha n^{n/2} (n-1) e^{-n/2}}{2^{n/2} \sqrt{\pi} \Gamma(\frac{n}{2} + \frac{1}{2})} \left(\frac{1}{\sigma_m} - \frac{1}{\sigma_M} \right) + \frac{\alpha \sqrt{n}}{\sqrt{2\pi}} \left(\frac{I_n}{\sigma_m} + \frac{1 - I_n}{\sigma_M} \right) + \frac{n^{n/2+1} e^{-n/2}}{2^{n/2} \Gamma(\frac{n}{2} + 1)} \ln \frac{\sigma_M}{\sigma_m} + 1,$$

where $I_1 \stackrel{\text{def}}{=} 1$ and

$$I_n \stackrel{\text{def}}{=} \frac{\sqrt{n}}{(\sqrt{2\pi})^{n-1}} \int_{\sum_{i=1}^{n-1} z_i^2 + (\sum_{j=1}^{n-1} z_j)^2 \leq n} \exp \left(-\frac{\sum_{i=1}^{n-1} z_i^2 + \left(\sum_{i=1}^{n-1} z_i \right)^2}{2} \right) d\mathbf{z}. \quad \square$$

3 Special cases

Constant mean, variance between σ_m^2 and σ_M^2 Using Theorem 5, we can also obtain the attenuation when the mean is fixed and the variance is allowed to vary. Let G_{σ_m, σ_M} the set of all Gaussian distributions over \mathbb{R} with mean 0 and whose variance lies between σ_m and σ_M . As before, we denote the collection of all pdfs on \mathbb{R}^n obtained by *i.i.d.* sampling from a distribution in G_{σ_m, σ_M} by G_{σ_m, σ_M}^n .

Corollary 6. For all $n \geq 1$

$$\hat{A}(G_{\sigma_m, \sigma_M}^n) = \frac{n^{n/2+1}e^{-n/2}}{2^{n/2}\Gamma(\frac{n}{2}+1)} \ln \frac{\sigma_M}{\sigma_m} + 1 = \sqrt{\frac{n}{\pi}} \ln \frac{\sigma_M}{\sigma_m} + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

Proof We obtain the above by setting $\alpha = 0$ in Theorem 5. The approximate value comes from Theorem 4. \square

It is easy to see that in the above result, the exact value of the mean does not matter so long as all Gaussians have the same mean. Therefore the attenuation of the collection of all Gaussians whose mean is β and whose variance is in between σ_m and σ_M remains the same as Corollary 6, namely

$$\frac{n^{n/2+1}e^{-n/2}}{2^{n/2}\Gamma(\frac{n}{2}+1)} \ln \frac{\sigma_M}{\sigma_m} + 1,$$

which is in turn equal to

$$\sqrt{\frac{n}{\pi}} \ln \frac{\sigma_M}{\sigma_m} + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

Constant variance σ^2 , mean between $-\alpha/2$ and $\alpha/2$ We have considered the case $\sigma^2 = 1$ in Theorem 1 already. Note that one could obtain Theorem 1 from Theorem 5 by setting $\sigma_m = \sigma_M = 1$. For any fixed $\sigma^2 > 0$ and all $n \geq 1$, the attenuation of length- n *i.i.d.* strings from the collection of all Gaussians with variance σ^2 and mean between $-\alpha/2$ to $\alpha/2$ is

$$1 + \frac{\alpha}{\sigma} \sqrt{\frac{n}{2\pi}}$$

by setting $\sigma_m = \sigma_M = \sigma$ in Theorem 5.

4 Acknowledgments

Narayana Santhanam was supported in this research by National Science Foundation Grants CCF-1065632 and CCF-1018984.

5 Author contributions

Both authors are equal contributors to this work. Both have read the manuscript and have approved it.

References

- [1] C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379—423, 623—656, 1948.
- [2] Y.M. Shtarkov. Universal sequential coding of single messages. *Problems of Information Transmission*, 23(3):3—17, 1987.
- [3] B. Fittingoff. Universal methods of coding for the case of unknown statistics. In *Proceedings of the 5th Symposium on Information Theory*, pages 129—135. Moscow-Gorky, 1972.
- [4] W. Feller. *An introduction to probability theory and its applications*. Wiley Series in Probability and Mathematical Statistics, 1966.
- [5] Ken iti Sato. *Levy Processes and infinitely divisible distributions*. Cambridge University Press, 1999.
- [6] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14(3):1080—1100, 1986.
- [7] R.E. Krichevsky and V.K. Trofimov. The preformance of universal coding. *IEEE Transactions on Information Theory*, 27(2):199—207, March 1981.
- [8] Q. Xie and A.R. Barron. Asymptotic minimax regret for data compression, gambling and prediction. *IEEE Transactions on Information Theory*, 46(2):431—445, March 2000.
- [9] J. Kieffer and E. Yang. Grammar based codes: A new class of universal lossless source codes. *IEEE Transactions on Information Theory*, 46(3):737—754, May 2000.
- [10] M. Drmota and W. Szpankowski. The precise minimax redundancy. In *Proceedings of IEEE Symposium on Information Theory*, 2002.
- [11] O. Kallenberg. *Foundations of modern probability*. Springer-Verlag, 1997.
- [12] Terrance Tao. The mesoscopic structure of gue eigenvalues. Available from <http://terrytao.wordpress.com/2010/12/17/the-mesoscopic-structure-of-gue-eigenvalues/>, Dec 2010. The article goes over several proofs of the Sylvester identity.
- [13] Keith Ball. An elementary introduction to modern convex geometry. In *Flavors of Geometry*, pages 1—58. Univ. Press, 1997.